Chemical Technology, Control and Management

Volume 2020 | Issue 4 Article 9

8-29-2020

AN EFFECTIVE METHOD FOR ATTRIBUTE SUBSET SELECTION, CONSIDERING THE RESOURCE IN PATTERN RECOGNITION

Bakhtiyorjon Bakirovich Akbaraliev

Tashkent University of Information Technologies named after Muhammad al-Khwarizimi Address: Amir Temur st., 1002000, Tashkent city, Republic of Uzbekistan E-mail:b.akbaraliev@gmail.com; b.akbaraliev@tuit.uz, Phone:+998-93-376-54-00., b.akbaraliev@gmail.com

ISSN 1815-4840, E-ISSN 2181-1105
Himičeskaâ tehnologiâ. Kontrol′ i upravlenie
CHEMICAL TECHNOLOGY. CONTROL AND MANAGEMENT
2020, №4 (94) pp.57-64
International scientific and technical journal
journal homepage: https://uzjournals.edu.uz/ijctcm/
Since 2005

UDC 004.8, 519.254

AN EFFECTIVE METHOD FOR ATTRIBUTE SUBSET SELECTION, CONSIDERING THE RESOURCE IN PATTERN RECOGNITION

Bakhtiyorjon Bakirovich Akbaraliev

Tashkent University of Information Technologies named after Muhammad al-Khwarizimi
Address: Amir Temur st., 1002000, Tashkent city, Republic of Uzbekistan
E-mail:b.akbaraliev@gmail.com; b.akbaraliev@tuit.uz, Phone:+998-93-376-54-00.

Abstract: *An analytical method for determining informative sets of features (INP) is developed, taking into account the resource for criteria based on the use of a measure of dispersion of classified objects. The areas of existence of the solution are defined. The statements and properties for the Fischer-type information criterion are proved, using which the proposed analytical method for determining the INP guarantees optimal results in the sense of maximizing the selected functional. The appropriateness of choosing this type of informative criterion is justified. A method for transforming attributes is proposed. The universality of the method in relation to the type of features is shown. An algorithm for implementing this method is given. In addition, the paper discusses the dynamics of the growth of information volume in the world, problems related to big data, as well as problems and tasks of pre-processing data. The relevance of reducing the dimension of the feature space for performing data processing and visualization without unnecessary difficulties is proved. The disadvantages of existing methods and algorithms for selecting an informative set of features are shown.*

Keywords: *training tuples (samples), attribute (feature), pattern recognition, informative attribute set, informativeness criterion, Big Data, dimensionality reduction, the scattering measure, informative vector, attribute (feature) space.*

Аннотация: *Таснифланувчи объектларнинг тарқоқлик ўлчовларига асосланган мезондан фойдаланган ҳолда, ресурсга боғлиқ белгиларнинг ахборотли тўплами (БАТ) аниқлаш бўйича аналитик усули ишлаб чиқилган. Қўйилган масалани ечимлари мавжуд бўладиган соҳаси аниқланган. Таклиф этилаётган аналитик усул орқали танланган БАТ, ахборотлилик мезони сифатида фойдаланилган, фишер кўринишидаги функционални максимумга эришиши оптимал натижа бериши тасдиқ ва хоссалар орқали исботланган. Бундай кўринишдаги ахборотлилик мезонини танланганлиги асосланган. Белгиларни трансформация қилиш бўйича ёндашув таклиф этилган. Усулни белгиларни турига нисбатан универсал бўлиши кўрсатилган. Мазкур усулни амалга ошириш алгоритми келтириб ўтилган. Бундан ташқари, ишда жаҳондаги ахборотларни ўсиш динамикаси, катта ҳажмдаги маълумотлар билан боғлиқ муаммолар ҳамда маълумотларга дастлабки ишлов бериш муаммо ва вазифалари муҳокама қилинган. Маълумотларни ортиқча қийинчиликларсиз қайта ишлаш ва визуаллаштириш учун белгилар фазосини пасайтириш долзарблиги асосланган. Белгиларнинг ахборотли тўпламини танлаш бўйича мавжуд усул ва алгоритмларнинг камчиликлари кўрсатиб ўтилган.*

Таянч сўзлар: *ўқув танланмаси, белги, тимсолларни таниб олиш, белгиларнинг ахборотли тўплами, ахборотлилик мезони, катта ҳажмли маълумотлар, ўлчамлиликни камайтириш, тарқоқлик ўлчови, ахборотли вектор, белгилар фазоси.*

Аннотация: *Разработан аналитический метод определения информативных наборов признаков (ИНП) с учетом ресурса для критериев, основанных на использовании меры рассеивания классифицируемых объектов. Определены области существования решения. Доказаны утверждения и свойства для критерия информативности фишеровского типа, с использованием которых предложенный аналитический метод определения ИНП гарантирует оптимальность результатов в смысле максимизации выбранного функционала. Обоснована уместность выбора такого вида критерия информативности. Предложен способ трансформации атрибутов. Показана универсальность метода в отношении типа признаков. Приведен алгоритм реализации данного метода. Кроме того, в работе обсуждены динамика роста объема информации в мире, проблемы, связанные с большими данными, а также проблемы и задачи предварительной обработки данных. Обоснована актуальность снижения размерности признакового пространства для осуществления без лишних затруднений*

обработки данных и их визуализации. Показаны недостатки существующих методов и алгоритмов выбора информативного набора признаков.

Ключевые слова: *обучающая выборка, признак, распознавание образов, информативный набор признаков, критерий информативности, большие данные, снижение размерности, меры рассеивания, информативный вектор, признаковое пространство.*

Введение

В последние годы объем информации растёт довольно высокими темпами. По данным международной исследовательской и консалтинговой компании, занимающейся изучением мирового рынка информационных технологий и телекоммуникаций (IDC), объём информации на электронных носителях в 2018 году был увеличен более чем на 700 эксабайт (10^{18} байт). Ожидается, что к 2023 году совокупный объём всех мировых хранилищ данных достигнет 11,7 зеттабайт (10^{21}байт) [1].

Такие данные (Big Data) часто создаются путем интеграции различных источников данных, соответствующих различным стандартам. В таких случаях, к сожалению, у нас практически нет возможности анализировать, систематизировать и отслеживать все эти данные. При отсутствии таких возможностей, часто, доля «грязных» данных (такие как, неточные, неполные, дублированные, противоречивые, зашумленные, бесполезные) увеличиваются пропорционально объему данных.

Как правило, подобные проблемы решаются с помощью методов и алгоритмов интеллектуального анализа данных (Data Mining). Практическое применение методов Data Mining предполагает многоэтапную процедуру, где одним из ключевых этапов является предварительная обработка данных. На этом этапе осуществляются [2-4]:

- очистка данных (***Data cleaning***), т.е. исключение противоречий, пропусков, случайных выбросов и помех;
- интеграция данных (***Data integration***), т.е.объединение данных из нескольких возможных источников в одном хранилище;
- преобразование данных (***Data transformation***), т.е. агрегирование и сжатие данных, дискретизация атрибутов, сокращение размерности и т.д.

Современные массивы данных, к которым могут быть применены те или иные методы Data Mining, могут характеризоваться большим числом признаков, формирующих признаковое пространство большой размерности. Поэтому актуальна задача снижения размерности такого пространства до размерности, позволяющей без лишних затруднений осуществлять обработку данных и (или) их визуализацию.

К настоящему времени разработаны и исследованы разнообразные подходы, методы и алгоритмы по снижению размерности признаков [5-35]. Все эти подходы могут быть разделены на два больших класса. Первый класс предусматривает трансформацию признакового пространства. Один из наиболее известных и применяемых на практике подходов этого класса – метод главных компонентов. Другой класс методов заключается в выборе наиболее информативных, полезных признаков и исключений из рассмотрения неинформативных признаков без трансформации исходного пространства. Здесь применяют методы и подходы: полного или усеченного перебора; ветвей и границ; эволюционные; со случайным выбором.

Анализ показывает, что многие существующие методы и алгоритмы определения информативного набора признаков (ИНП) представляют собой тот или иной способ частичного перебора и поэтому не гарантируют достижения оптимального результата в смысле используемого критерия информативности. Более того, в этих методах не учтены многие факторы, связанные с затратами определения признаков, кроме вычислительных затрат. В связи с чем целесообразна разработка новых методов, которые учитывали бы другие затраты и были бы основаны на использовании развитого аппарата математического программирования, давала оптимальное (или близкое к нему) решение.

Цель данной работы состоит в разработке аналитического метода определения ИНП с учетом ресурса для критериев, основанных на использовании меры рассеяния классифицируемых объектов.

Методы исследования и полученные результаты

Пусть задана обучающая выборка объектов X:

$$X = \bigcup_{p=1}^{r} X_p,\ X_p \cap X_q = \emptyset, (p \neq q,\quad p,q = \overline{1,r}), X_p = \{x_{pi}: i = \overline{1,m_p}\}, \tag{1}$$

где X_p $(p = \overline{1,r})$ – классы, – количество классов, x_{pi} – объект p– го класса, m_p– количество объектов p –го класса.

Тогда общее количество объектов обучающей выборки равно:

$$M = card(X) = \sum_p card(X_p) = \sum_p m_p.$$

Предположим, каждый объект обучающей выборки является мерным вектором признаков, тогда $\forall x_{pi} = (x_{pi}^1, x_{pi}^2, \dots, x_{pi}^N) \in X_p$и $\dim(X) = N$.

Для снижения размерности исходного пространства признаков и выбора ИНП будем использовать -мерный вектор $\lambda = (\lambda^1, \lambda^2, \dots, \lambda^N)$, т.е.

$$\lambda: X \to X|_\lambda = \{x|\ x = (\lambda^1 x^1,\ \lambda^2 x^2, \dots, \lambda^N x^N)\}, \tag{2}$$

где $\lambda^j \in \{0,1\}$ $(j = \overline{1,N})$.

Здесь $\lambda^j = 0$ указывает на отсутствие или $\lambda^j = 1$ наличие j-го признака в рассматриваемом наборе.

Определение 1[*]. Вектор $\lambda = (\lambda^1, \lambda^2, \dots, \lambda^N)$ называется ℓ информативным, если $\sum_{j=1}^{N} \lambda^j = \ell$.

Множество всех ℓ информативных векторов обозначим через Λ^ℓ:

$$\Lambda^\ell = \left\{ \lambda|\ \sum_{j=1}^{N} \lambda^j = \ell\ ,\ \lambda^j \in \{0,1\}, j = \overline{1,N} \right\}. \tag{3}$$

Отсюда видно, что мощность множества Λ^ℓ равна

$$|\Lambda^\ell| = C_N^\ell = \frac{N!}{\ell!(N-\ell)!}. \tag{4}$$

Из (2) и (3) следует:

$$\Lambda^\ell: X \to X|_{\Lambda^\ell} = \{X|_\lambda \mid\ \lambda \in \Lambda^\ell\}. \tag{5}$$

Принимая во внимания, что задача определения оптимального набора признаков обычно связана с оценкой качества классификации, для отбора оптимального ИНП из (5) будем использовать функционал $I(\lambda)$ в качестве критерия информативности (эффективности).

Определение 2. Назовем подсистему (или ИНП) $X|_\lambda \in X_{\Lambda^\ell}$ оптимальной, если $\exists \lambda \in \Lambda^\ell$ для которого верно $I(\lambda) = \text{extr}_{\mu \in \Lambda^\ell} I(\mu)$.

Тогда задачу определения информативного набора признаков можно свести к оптимизационной задаче

$$\begin{cases} I(\lambda) \to extr \\ \lambda \in \Lambda^\ell \end{cases}. \tag{6}$$

Предположим, что каждый признак объекта требует определенных затрат (технических, вычислительных, временных и т.д.) и для определения объекта выделен фиксированный ресурс. Для обозначения вышеизложенного введем вектор $c = (c^1,\ c^2, \dots, c^N)$, где c^j – указывает общие затраты для определения признака -го признака $(j = \overline{1,N})$. Пусть C_0 - фиксированный ресурс, выделенный для определения объекта.

Если учитывать, что каждый $\lambda \in \Lambda^\ell$ вектор однозначно определяет конкретную ИНП, то (c, λ) будет показывать, сколько требуется ресурсов, чтобы получить этот ИНП.

Можем ли получить этот набор признаков? Если $(c, \lambda) \leq c_0$, то да.

Теперь можно сформулировать общую математическую постановку задачи выбора набора наиболее информативных признаков при ограниченности ресурсов следующим образом:

$$\begin{cases} I(\lambda) \to extr, \\ \lambda \in \Lambda^{\ell}, \\ (\mathrm{c}, \lambda) \le c_{0,} \end{cases} \tag{7}$$

где $(*,*)$ – скалярное произведение векторов.

Если конкретно неизвестно об объектах и классах, кроме значения признаков объекта, то можно предполагать следующее.

Гипотеза 1. Если обучающая выборка имеет вид (1), то значение признаков между объектами одного и того же класса более схожие (близкие), чем значение разных классов.

Гипотеза 2. Если гипотеза 1 верна, то оптимальный ИНП сближает объекты одного класса и разделяет объекты разных классов лучше, чем другие.

Опираясь на вышеизложенные гипотезы критерия информативности (эффективности) для выбора ИНП определим в виде функционала следующим образом [9]:

$$I(\lambda) = \frac{(a,\lambda)}{(b,\lambda)}, \tag{8}$$

где $I(\lambda)$ – однородный функционал фишерского типа, $a = (a^1, a^2, \dots, a^N)$ – межклассовое расстояние, $b = (b^1, b^2, \dots, b^N)$ – внутриклассовое расстояние.

Коэффициенты a^j, b^j не зависят от λ и вычисляются заранее.

Для данного функционала считается, что ИНП является оптимальным, если значения функционала больше. Среди основных достоинств функционала следует выделить его относительную простоту.

С другой стороны, простота Функционала Фишера выступает в качестве недостатка, поскольку здесь можно "упустить из виду" сложные нелинейные свойства анализируемых классов. Однако, в пользу данного функционала служит тот факт, что простые критерии качества, как правило, оказываются более надежными, т.е. выделяют если не самую информативную, то, по крайней мере, достаточно информативную подсистему признаков. И наоборот, сложные критерии, позволяют в большинстве случаев более информативные подсистемы позволяют все же выбрать подсистемы, для которых трудно построить решающее правило.

Тогда (7) будет имеет следующий вид

$$\begin{cases} I(\lambda) = \frac{(a,\lambda)}{(b,\lambda)} \to max, \\ \lambda \in \Lambda^{\ell}, \\ (\mathrm{c}, \lambda) \le c_0. \end{cases} \tag{9}$$

Прежде, чем приступить к решению (9), сначала определим область существования решения. Будем анализировать ресурс вектор $c = (c^1, c^2, \dots, c^N)$. Существует такая последовательность попарно различных индексов $j_1, j_2, \dots, j_N$ – таких, при которых

$$c^{j_1} \le c^{j_2} \le \dots \le c^{j_N} \tag{10}$$

Построим числовые последовательности, учитывая (10), следующим образом

$$f_1 = \sum_{i=1}^{\ell} c^{j_i}, f_2 = \sum_{i=2}^{\ell+1} c^{j_i}, \dots, f_k = \sum_{i=k}^{\ell+k-1} c^{j_i}, \dots, f_{N-\ell+1} = \sum_{i=N-\ell+1}^{N} c^{j_i} \tag{11}$$

Утверждение 1. Из числовой последовательности (11)вытекают следующее:

1) если$f_1 > c_0$, то (9) не имеет решения;
2) если $\mathrm{f}_1 \le \mathrm{c}_0$, то (9) имеет хотя бы одно решение;
3) если$\exists \mathrm{t} > \ell$ – такое, что $\mathrm{f}_1 - \mathrm{c}^{\mathrm{j}_1} + \mathrm{c}^{\mathrm{j}_t} \le \mathrm{c}_0 < \mathrm{f}_1 - \mathrm{c}^{\mathrm{j}_1} + \mathrm{c}^{\mathrm{j}_{t+1}}$, то признаки объектов, соответствующие $\mathrm{j}_{t+1}, \mathrm{j}_{t+2}, \dots, \mathrm{j}_N$ индексам, не будут участвовать в ИНП и должны исключаться из дальнейшего рассмотрения, т.е. эти признаки должны быть исключены из признакового пространства.

Доказательство.

1. Пусть $f_1 > c_0$.Тогда $\forall \lambda \in \Lambda^{\ell}$ вектора$(\mathrm{c}, \lambda) \ge c_0$. Значит, (9) не имеет решения■

2. Пусть $f_1 \le c_0$.Рассмотрим вектор $\lambda = (\lambda^1, \lambda^2, \dots, \lambda^N)$,

где $\lambda^j = \begin{cases} 1, \text{если} j \in \{j_1, j_2, \dots, j_N\} \\ 0, \text{ в противном случае} \end{cases}$.Тогда $\lambda \in \Lambda^\ell$и$(c, \lambda) \le c_0$. Значит (9) имеет решения■

3. Пусть выполнены условия 3) Утверждения 1. Если $\lambda \in \Lambda^\ell$ и$\sum_{i=t+1}^{N} \lambda^{j_i} > 0$, то $(c, \lambda) \ge c_0$■

Рассмотрим следующую оптимизационную задачу

$$\begin{cases} I(\lambda) = \frac{(a,\lambda)}{(b,\lambda)} * s(\lambda) \to max, \\ \lambda \in \Lambda^\ell, \end{cases} \quad (12)$$

где $s(\lambda) = \begin{cases} -1, & c_0 - (c, \lambda) < 0 \\ 1, & c_0 - (c, \lambda) \ge 0 \end{cases}$.

По аналогии [***] для решения задачи (12) введем вектор-функцию

$$\varphi(\lambda) = a(b, \lambda) - b(a, \lambda), \quad (13)$$

которая указывает направление наискорейшего роста функционала $I(\lambda)$ в точке λ.

Пусть $b^j > 0$ $(j = \overline{1, N})$и$\lambda, \mu \in \Lambda^\ell$.

Определим связи между (12) и (13).

Пусть $(\varphi(\lambda), \mu) \ge 0$.Тогда

$(a(b, \lambda) - b(a, \lambda), \mu) \ge 0 \Longrightarrow (a, \mu)(b, \lambda) - (b, \mu)(a, \lambda) \ge 0 \Longrightarrow$

$$\Longrightarrow (a, \mu)(b, \lambda) \ge (b, \mu)(a, \lambda) \Longrightarrow \frac{(a,\mu)}{(b,\mu)} \ge \frac{(a,\lambda)}{(b,\lambda)}. \quad (14)$$

Отсюда вытекает следующее

Свойство 1.Если $(\varphi(\lambda), \mu) \ge 0$, то

a) $I(\mu) \le I(\lambda) \le 0$при$\boldsymbol{s(\lambda) = -1, s(\mu) = -1}$;
b) $I(\mu) \ge -I(\lambda) \ge 0$при$\boldsymbol{s(\lambda) = -1, s(\mu) = 1}$;
c) $I(\mu) \le -I(\lambda) \le 0$при$\boldsymbol{s(\lambda) = 1, s(\mu) = -1}$;
d) $I(\mu) \ge I(\lambda) \ge 0$при$\boldsymbol{s(\lambda) = 1, s(\mu) = 1}$.

Пусть $I(\mu) \ge I(\lambda)$, тогда

$$\frac{(a,\mu)}{(b,\mu)} s(\mu) \ge \frac{(a,\lambda)}{(b,\lambda)} s(\lambda) \Longrightarrow (a, \mu)(b, \lambda)s(\mu) - (a, \lambda)(b, \mu)s(\lambda) \ge 0 \Longrightarrow$$

$$\Longrightarrow (a(b, \lambda)s(\mu) - b(a, \lambda)s(\lambda), \mu) \ge 0. \quad (15)$$

Отсюда вытекает следующее

Свойство 2. Если $I(\mu) \ge I(\lambda)$, то

a) $(\varphi(\lambda), \mu) \le 0$при$\boldsymbol{s(\lambda) = -1, s(\mu) = -1}$;
b) $(a(b, \lambda) + b(a, \lambda), \mu) \ge 0$при$\boldsymbol{s(\lambda) = -1, s(\mu) = 1}$;
c) $(a(b, \lambda) + b(a, \lambda), \mu) \le 0$при$\boldsymbol{s(\lambda) = 1, s(\mu) = -1}$;
d) $(\varphi(\lambda), \mu) \ge 0$при$\boldsymbol{s(\lambda) = 1, s(\mu) = 1}$.

Утверждение 2. Если $s(\lambda) = s(\mu) = 1$,то $I(\mu) \ge I(\lambda)$ тогда и только тогда, когда $(\varphi(\lambda), \mu) \ge 0$.

Доказательство.

$\Longrightarrow$)вытекает из d) свойство 2.

$\Longleftarrow$)вытекает из d) свойство 1■

Для компонентов вектор-функции $\varphi(\lambda)$ существует такая последовательность попарно различных индексов $\exists\, t_1, t_2, \dots, t_N$ – таких, при которых $\varphi(\lambda)^{t_1} \ge \varphi(\lambda)^{t_2} \ge \dots \ge \varphi(\lambda)^{t_N}$.

Утверждение 3. Если для вектора $\mu(\lambda)$:$\mu^{t_1} = \mu^{t_2} = \dots = \mu^{t_\ell} = 1$ и $\mu^{t_{\ell+1}} = \dots = \mu^{t_N} = 0$, то $\mu(\lambda) \in \Lambda^\ell$и$(\varphi(\lambda), \mu(\lambda)) = max\{(\varphi(\lambda), \eta) \mid \eta \in \Lambda^\ell\}$.

Утверждение 4. $\forall \lambda \in \Lambda^\ell$верно $\big(\varphi(\lambda), \mu(\lambda)\big) \ge 0$.

Доказательство.

$\big(\varphi(\lambda), \mu(\lambda)\big) = (a(b, \lambda) - b(a, \lambda), \mu(\lambda)) \ge (a(b, \lambda) - b(a, \lambda), \lambda) = (a, \lambda)(b, \lambda) - (b, \lambda)(a, \lambda) = 0$. ■

Пусть существуют решения для задачи (9), тогда имеет место следующие:

Утверждение 5. $\forall\lambda \in \Lambda^{\ell}$существует $\exists s\big(\mu(\lambda)\big) = 1$ и$I(\mu(\lambda)) \geq I(\lambda)$.

Теорема. Если $s\big(\mu(\lambda)\big) = 1$и$I(\lambda) = I(\mu(\lambda))$, то $I(\lambda) = max\{I(\eta) \mid \eta \in \Lambda^{\ell}\}$.

Доказательство.

Пусть $s\big(\mu(\lambda)\big) = 1$и$I(\lambda) = I(\mu(\lambda))$. Тогда из свойства 2 вытекает, что $s(\lambda) = 1$. Из утверждения 2 вытекает $\big(\varphi(\lambda), \mu(\lambda)\big) = 0$. Отсюда, согласно утверждению 3

$0 = (\varphi(\lambda), \mu(\lambda)) = max\{(\varphi(\lambda), \eta) \mid \eta \in \Lambda^{\ell}\} \Longrightarrow (\varphi(\lambda), \eta) \leq 0$для $\forall\eta \in \Lambda^{\ell} \Longrightarrow I(\lambda) \leq I(\eta)$. ■

Заметим, что теорема гарантирует оптимальность полученного решения, т.е. значение функционала $I(\lambda)$ при найденном решении λ достигает своего максимума на множестве Λ^{ℓ}.

Итак, предлагаемый метод состоит в следующем. На первом шаге выбирается произвольный $\lambda \in \Lambda^{\ell}$ вектор, например,

$\lambda = (\underbrace{1, 1, \dots, 1}_{\ell}, 0, \dots, 0)$

В качестве $\mu(\lambda)$берем

$$\max\{\eta(\lambda) \mid \big(\varphi(\lambda), \eta(\lambda)\big) \geq 0,\ s\big(\eta(\lambda)\big) = 1,\ \eta(\lambda) \in \Lambda^{\ell}\}. \qquad (16)$$

Далее на каждом шаге вектор λ заменяется на $\mu(\lambda)$, удовлетворяющий условия (16), до тех пор, пока функционал $I(\lambda)$ растет, т.е. пока не выполнены условия $I(\lambda) = I(\mu(\lambda))$. Как только функционал перестанет расти, вычисление заканчивается.

Сформулируем рассматриваемый метод в виде алгоритма:

Шаг 1. Входными параметрами являются значения: ℓ- требуемое число признаков; N - количество признаков; c, b, c -N мерные векторы; c_0 – выделенный ресурс.

Шаг 2. Задание вектора: $\lambda = (\underbrace{1, 1, \dots, 1}_{\ell}, 0, \dots, 0)$;

Шаг 3. Вычисление:$I(\lambda)$;

Шаг 4. Определение:$\mu(\lambda) = \max\{\eta(\lambda) \mid \big(\varphi(\lambda), \eta(\lambda)\big) \geq 0,\ s\big(\eta(\lambda)\big) = 1,\ \eta(\lambda) \in \Lambda^{\ell}\}$;

Шаг 5. Вычисление:$I(\mu(\lambda))$.

Шаг 6. Сравнение:$I\big(\mu(\lambda)\big) > I(\lambda)$. Если неравенство выполняется, то полагаем $\lambda = \mu(\lambda)$,$I(\lambda) = I\big(\mu(\lambda)\big)$и переходим к шагу 4. В противном случае процедура заканчивается и λ - оптимальное решение.

Шаг 7. Выходные параметры: λ,$I(\lambda)$.

Далее приводится способ определения меры схожести (близости) или меры рассеяния для объектов обучающей выборки.

Пусть признаки объектов принимают бинарные (т.е. $\{0,1\}$) и/или непрерывные (на отрезке $[\alpha, \beta]$) значения. Рассмотрим оба случая подробнее.

Как отмечено выше, $a = (a^1, a^2, \dots, a^N)$ – межклассовое расстояние и $b = (b^1, b^2, \dots, b^N)$ – внутриклассовое расстояние.

1) Пусть j_k-й ($j_k \in \{1,2, \dots, N\}$) признак принимает непрерывные значения на отрезке, т.е. $x^{j_k} \in [\alpha_{j_k}, \beta_{j_k}] \subset R$ (здесь R – множество действительных чисел). Для таких признаков в качестве меры рассеяния можно использовать Евклидову метрику. Тогда a^{j_k} и b^{j_k} определяется следующим образом:

$$a^{j_k} = \sum_{p,q=1}^{r}\big(\bar{x}_p^j - \bar{x}_p^j\big)^2,$$

$$b^{j_k} = \sum_{p=1}^{r}\left[\frac{1}{m_p}\sum_{i=1}^{m_p}\big(\bar{x}_p^j - x_{pi}^j\big)^2\right],$$

$$\bar{x}_p^{j_k} = \frac{1}{m_p}\sum_{i=1}^{m_p} x_{pi}^{j_k},$$

где $\bar{x}_p = \big(\bar{x}_p^1, \bar{x}_p^2, \dots, \bar{x}_p^N\big)$ – усреднённый объект класса $X_p(p = \overline{1, r})$.

2) Пусть $t_k \in \{1,2,\dots,N\}$ признак принимает бинарные значения, т.е.$x^{t_k} \in \{0,1\}$. Тогда a^{t_k} и b^{t_k} можно определить следующим образом:

$$a^{t_k} = \sum_{p,q=1}^{r} \left[\left| \frac{m_p^{t_k}(0)}{m_p} - \frac{m_q^{t_k}(1)}{m_q} \right| * \left| \frac{m_p^{t_k}(1)}{m_p} - \frac{m_q^{t_k}(0)}{m_q} \right| \right],$$

$$b^{t_k} = \sum_{p=1}^{r} \frac{m_p^{t_k}(0) * m_p^{t_k}(1)}{m_p},$$

где $m_p^{t_k}(i)$ – количество объектов класса X_p $(p = \overline{1,r})$, которые принимают значения i.

Заключение

Основным результатом данной работы является решение научной проблемы выбора информативных наборов признаков с учетом ресурса в задачах распознавания образов и предварительной обработки данных. В ходе решения этой проблемы осуществлено следующее:

- изучены основные причины и типы проблем больших данных;
- приведены основные задачи предварительной обработки данных;
- показана необходимость снижения размерности признакового пространства и недостатки существующих методов по выбору ИНП;
- разработан и исследован эффективный аналитический метод определения ИНП с учетом ресурса;
- предложен подход определения области существования решения для задачи выбора ИНП при ограниченности ресурса;
- доказано, что предложенный метод определения ИНП гарантирует оптимальность результатов в смысле максимизации выбранного функционала;
- обоснована уместность выбора такого вида критерия информативности;
- показана универсальность метода в отношении типа признаков;
- приведен алгоритм реализации данного метода.

References:

1. https://regnum.ru/news/it/2574265.html (data obrasheniya: 15.01.2020)
2. Jiawei Han, Micheline Kamber, Jian Pei. Data mining : concepts and techniques// 3rd ed. by Elsevier Inc., USA, 2012.
3. Zamyatin A.V. Intellektualniy analiz dannix//Tomsk : Izd.dom TomGU, 2016.
4. Jian Long Zhou, Fang Chen. Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent//Springer, Human-Computer interaction Series, 2018, Switzerland, p.482.
5. Zagoruyko N.G. Prikladnie metodi analiza dannix i znaniy//Novosibirsk: IM SO RAN, 1999, str. 270.
6. Nasma M. Sovremennie tendensii metodov intellektualnogo analiza dannix: metod klasterizatsii//Moskovskiy ekonomicheskiy jurnal, №6, Rossiya, 2019.
7. Nishanov A.X., Akbaraliev B.B., Ruzibaev O.B., Xujaev O.K. Sravnitelniy analiz algoritmov na osnove nechetkogo K-srednix s primeneniem razlichnix metrik // Kimyoviy texnologiya, nazorat va boshqaruv, Xalqaro ilmiy-texnikaviy jurnal, 2014 yil, 6-son, 78-82 b.
8. Kamilov M.M., Nishanov A.H., Akbaraliev B.B. About one clustering algorithm in intellectual data analysis// Proceedings of ICEIC2008, June 24-27, 2008, Tashkent, pp. 476-478.
9. Kamilov M.M., Nishanov A.H., Akbaraliev B.B. Methods of forming of optimal sign space for object recognition in the class of logic-heuristic algorithms// Fourth World Conference on Intelligent Systems for Industrial Automation - WCIS 2006, Tashkent.
10. Akbaraliev B.B. Formirovanie informativnix naborov priznakov v slojnix sistemax raspoznavaniya// TATU xabarlari, №2, 2007, 47-50 b.
11. Raxmanov A.T., Akbaraliev B.B., Ergashev A.K. Ob odnom metode sokrashenie razmernosti ob'ema viborki v intellektualnom analize dannix//"Informatika va Energetika muammolari" O'zbekiston jurnali, 1-son, 2011, 76-79 b.
12. Sunita Beniwal, Jitender Arora. Classification and Feature Selection Techniques in Data Mining//International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 1 Issue 6, August – 2012.
13. Huiqing Liu, Jinyan Li, Limsoon Wong. A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns//Genome Informatics 13: 51-60, 2002.
14. Krasnyanskiy M.N. i dr. Sravnitelniy analiz metodov mashinnogo obucheniya dlya resheniya zadachi klassifikatsii dokumentov nauchno-obrazovatelnogo uchrejdeniya // Vestnik VGU, seriya: Sistemniy analiz i informatsionnie texnologii, 2018, № 3, str.173-182.

15. Lbov G. S. Metodi obrabotki raznotipnix eksperimentalnix dannix // Novosibirsk: Nauka, Sib.otd., 1981. - 160 s.
16. Juravlev Yu.I. Izbrannie nauchnie trudi//M: Izdatelstvo Magistr, 1998. – 420s.
17. Zhang, L., Luo, M., Liu, J., Li, Z., Zheng, Q. Diverse fuzzy c-means for image clustering //Pattern Recognition LettersVolume 130, February 2020, Pages 275-283.
18. Santra, D., Basu, S.K., Mandal, J.K., Goswami, S. Rough set based lattice structure for knowledge representation in medical expert systems: Low back pain management case study//Expert Systems with Applications Volume 145, 1 May 2020, 113084
19. Xiong, Y., Zuo, R. Recognizing multivariate geochemical anomalies for mineral exploration by combining deep learning and one-class support vector machine//Computers and Geosciences Volume 140, July 2020, 104484
20. Gai, J., Shen, J., Wang, H., Hu, Y. A Parameter-Optimized DBN Using GOA and Its Application in Fault Diagnosis of Gearbox//Shock and Vibration, Volume 2020, 2020, 4294095.
21. Raja, P.S., Thangavel, K. Missing value imputation using unsupervised machine learning techniques//Soft Computing 24(6), c. 4361-4392,2020.
22. Wang,D., Tian,F., Yang,S.X., Jiang,D., Cai,B. Improved deep CNN with parameter initialization for data analysis of near-infrared spectroscopy sensors//Sensors (Switzerland) Volume 20, Issue 3, 1 February 2020, 874, 20(3),874, 2020.
23. Lou, P., Jimeno Yepes, A., Zhang, Z., Li, C., Wren, J. BioNorm: Deep learning-based event normalization for the curation of reaction databases//Bioinformatics Volume 36, Issue 2, 15 January 2020, Pages 611-620.
24. Fu, S., Liu, X. A new method to solve the problem of facing less learning samples in signal modulation recognition//Eurasip Journal on Wireless Communications and Networking Volume 2020, Issue 1, 1 December 2020, 8.
25. Wei, D., Chen, T., Li, S., Zhao, Y., Li, T. Adaptive dictionary learning based on local configuration pattern for face recognition//Eurasip Journal on Advances in Signal Processing Volume 2020, Issue 1, 1 December 2020, 20.
26. Ala'raj, M., Majdalawieh, M., Abbod, M.F. Improving binary classification using filtering based on k-NN proximity graphs//Journal of Big Data, Volume 7, Issue 1, 1 December 2020, 15.
27. Mishra, G., Vishwakarma, V.P., Aggarwal, A. Constrained L1-optimal sparse representation technique for face recognition/Optics and Laser Technology Volume 129, September 2020, 106232.
28. Kibbey, T.C.G., Jabrzemski, R., O'Carroll, D.M.Supervised machine learning for source allocation of per- and polyfluoroalkyl substances (PFAS) in environmental samples//Chemosphere Volume 252, August 2020, 126593.
29. Shen, Z., Man, Z., Cao, Z., Zheng, J. A new intelligent pattern classifier based on structured sparse representation //Computers and Electrical Engineering Volume 84, June 2020, 106641.
30. Nishanov A.Kh., Djurayev G.P., Kasanova M.Kh. Improved algorithms for calculating evaluations in processing medical data // National Institute of Science Communication and Information Resources (NISCAIR)-India, 2019,-3158-3165.
31. Kamilov M., Nishanov A., Beglerbekov R. Modified stages of algorithms for computing estimates in the space of informative features // International Journal of Innovative Technology and Exploring Engineering (2019) 8(6).
32. Nishanov A. Avazov E. Akbaraliyev B. Partial selection method and algorithm for determining graph-based traffic routes in a real-time environment// International Journal of Innovative Technology and Exploring Engineering (2019) 8(6) 696-698 ISSN: 22783075.
33. Emary E. Zawbaa H. Hassanien A. Binary ant lion approaches for feature selection// Neurocomputing. 2016 vol: 213. DOI 10.1016/j.neucom.2016.03.101. ISSN 18728286.
34. Yong Z. Dun-wei G. Wan-qiu Z. Feature selection of unreliable data using an improved multi-objective PSO algorithm// Neurocomputing. 2016 vol: 171. DOI 10.1016/j.neucom.2015.07.057. ISSN 18728286.
35. Zhang Y. Gong D. Sun X. Guo Y. A. PSO-based multi-objective multi-label feature selection method in classification.// Scientific Reports. 2017 vol: 7 (1). DOI 10.1038/s41598-017-00416-0. ISSN 20452322.